Journal of Business Research.

Title:

Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type

Authors and Affiliations

Joni Salminen[1], Vignesh Yoganathan[2] (corresponsing author), Juan Corporan[3], Bernard J. Jansen[4], and Soon-Gyo Jung[4]

[1] *Qatar Computing Research Institute, Hamad Bin Khalifa University*
Qatar; and
Turku School of Economics at the University of Turku
Finland

[2] *University of Bradford, School of Management*
Emm Lane, Bradford BD9 4JL
United Kingdom

[3] *Banco Santa Cruz RD*
Dominican Republic

[4] *Qatar Computing Research Institute, Hamad Bin Khalifa University*
Qatar

Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type

Abstract

As complex data becomes the norm, greater understanding of machine learning (ML) applications is needed for content marketers. Unstructured data, scattered across platforms in multiple forms, impedes performance and user experience. Automated classification offers a solution to this. We compare three state-of-the-art ML techniques for multilabel classification - Random Forest, K-Nearest Neighbor, and Neural Network - to automatically tag and classify online news articles. Neural Network performs the best, yielding an F1 Score of 70% and provides satisfactory cross-platform applicability on the same organisation's YouTube content. The developed model can automatically label 99.6% of the unlabelled website and 96.1% of the unlabelled YouTube content. Thus, we contribute to marketing literature via comparative evaluation of ML models for multilabel content classification, and cross-channel validation for a different type of content. Results suggest that organisations may optimise ML to auto-tag content across various platforms, opening avenues for aggregated analyses of content performance.

Keywords

machine learning; auto-tagging; web content; content marketing; neural network; digital marketing

1. Introduction

Turning online content into structured data is important for content marketers, as structuring the content supports users' information consumption and sharing purposes, and therefore, from a commercial perspective for firm performance (Balducci and Marinova, 2018). For marketers and decision-makers, especially in firms dealing with online content (e.g., social media managers, editors, content producers), a higher order understanding of content performance is crucial for competitive success, given the rising demand among users for personalised offerings (Kumar, 2018). Yet, making sense of online content performance to derive business value can be a daunting task, as the nature of data involved is complex in terms of volume and dynamics, it is fragmented across many channels, and it can be associated with many different metrics (Chun, 2018; Clarke and Jansen, 2017). Content classification (e.g. dividing the content into topics) is therefore a necessity, such that individual units of content are thematically aggregated to increase interpretability for decision-making in relation to content marketing[1] activities such as content creation, dissemination, and management. Nonetheless, beyond the obvious impracticalities of time and effort involved, manually tagging online content for keywords is problematic for two main reasons: a) the tagging process is fallible owing to human error; and b) classification taxonomies can change over time as new topics emerge, especially given the vast quantity of online data generated daily. Consequently, online content often remains largely unstructured with the absence or incorrect allocation of tags (Kutlu *et al.*, 2018). Machine learning approaches have emerged as a potential solution to this problem and are increasingly applied in a variety fields to uncover hidden insights by automating the classification process (Antons and Breidbach, 2018).

Even so, the application of machine learning approaches in *marketing* is still at a developmental stage, in need of refinement and insight (Balducci and Marinova, 2018; Sterne, 2018). In this research, we contribute to the marketing literature by: 1) Comparing three relevant approaches to automatically classify news articles based on web content from a major worldwide news and media organisation; 2) Developing and illustrating a neural network algorithm to address the multilabel classification issue in automatically classifying webpages containing news articles;

[1] We define content marketing as a strategic marketing action that consists of producing original digital and analogous multimedia content (e.g., text, video, pictures, infographics) whose goal is to entertain and inform consumers. The main difference between paid advertising and content marketing is that content marketing typically aims at organic dissemination of the content; i.e., instead of the firm paying for exposure, its followers actively share the content among their social networks.

and 3) Applying the same algorithm, without channel-specific training, on the same organisation's YouTube channel to test the generalisability of the approach. The latter evaluation is important for several reasons. Most notably, evaluation of the cross-channel applicability of automatic classification approaches is often not conducted in the research dealing with auto-tagging online content, which means that the generalisability of the models over time and in different channels is not properly addressed. Rather, researchers employing machine learning methods to this problem tend to utilize the test data from the same overall sample to evaluate their models' performance. Even though this practice is typical for evaluating a model's performance (i.e., machine learning models are tested such that training and test data are kept separate, so that the model does not "see" the test data prior to predicting it), the cross-sectional nature of data collection (i.e., the training and testing data belong to the same *overall* sample) makes it difficult to evaluate the model's true generalisability over time and in different channels. Therefore, by evaluating the cross-channel applicability of our model, we address the broader question: *Are machine learning models developed for online content classification generalisable beyond the dataset they were trained and tested on?* To address this question, we conduct a repeated test of the model on an independently collected dataset of the organisation's content, i.e., the titles and descriptions of the videos in the organisation's YouTube channel.

In addition to addressing a research gap within the automatic classification of online content, cross-channel applicability of tagging online content is highly important for organisations practically engaged in content marketing, as such organisations typically publish their content in multiple channels, including website and social media such as Facebook, Twitter, YouTube, and LinkedIn. Thus, when developing a classifier to tag the content published in different channels, the classifier needs to be able to perform well in a multichannel environment that the marketing mix of the modern content marketer consists of. With increasingly large, complex, and dynamic data becoming the basis of marketing decisions, it is ever more important to develop better methods of converting unstructured 'big' data into actionable information and insights (Syam and Sharma, 2018). Though the vast amount of available data is useful for training machine learning algorithms to make accurate predictions or classifications, developing the right approach can be challenging, not least because of the level of noise in the datasets and the diverse range of problems in relation to available technologies (Flake *et al.*, 2004). On the whole, higher level description of online content is important for machine-readability, model development, and statistically correlating topics to various key performance

metrics of content marketing such as visitor statistics, development of content coverage over time, or the range of topics covered by various websites. Our aim is to address the gap in the extant marketing literature for more advanced and innovative methods (Hofacker, 2012; Kumar, 2018) by comparing machine learning approaches to dealing with the multilabel classification problem when classifying news articles and examining a high-performing machine learning model's cross-channel applicability for a different type of content.

By using data from a worldwide news organisation, we show that our approach yields an overall F1 Score of 70%, even with a large set of topics. We further visualise the development of news articles over time; provided the taxonomy is updated with at least some examples, our classification is robust to topic changes and new topics emerging over time. In addition, we evaluate cross-platform applicability by classifying the same organisation's YouTube videos and then manually reviewing the results via three human coders.

The remainder of the paper is organised as follows. First, we present an overview of the literature on machine learning applications in marketing, followed by a summary of the proposed solution strategy. Next, we explain the data exploration and preparation procedure. We then evaluate three classifiers: Random Forests, K-Nearest Neighbors, and Neural Network (NN); followed by a more detailed application of NN whereby data collected from one year (2017) is used for training and data collected from another year (2018) is used for testing. Based on this, keywords are generated for unclassified news articles using the developed approach. Subsequently, we evaluate the cross-channel applicability by classifying YouTube videos of the news organization. Finally, we discuss implications and avenues for further research.

2. Machine learning in marketing and content classification

Machine learning is an umbrella term used to describe a variety of computer-based techniques for data mining to uncover complex patterns, particularly in large and complex datasets (Pereira *et al.*, 2018), with a view to deriving insights for prediction, classification, and decision-making purposes (Cui *et al.,* 2006). Particularly, in the context of a multiplicity of social media and user-generated content (UGC) platforms, the diversity of data, in both type and content, is as daunting a challenge as the volume of data that needs analysis. As a result, marketing research and applications are increasingly turning to the computational prowess of machine learning approaches (Syam and Sharma, 2018); as in the case of developing a highly optimised ranking system for hotels based on previous bookings, users' search engine behaviour and the content they generate on various social media platforms (see: Ghose *et al.,* 2012), or for auto-ranking

images (more unstructured than text) based on specific themes from the viewer's perspective to help bridge the projected vs. perceived image gap in destination-marketing (see: Deng and Li, 2018). From a statistical point of view, machine learning approaches are essentially not confined by limitations relating to linearity and the parametric nature of regular statistical analysis (Cui and Curry, 2005; Syam and Sharma, 2018).

Nascent studies in the marketing literature reveal some notable use of machine learning approaches to providing decision-support for problems in areas ranging from direct marketing (Cui and Wong, 2004; Ha *et al*., 2005) to strategic marketing (Martínez-López and Casillas, 2009; Orriols-Puig *et al.*, 2013). Among the various applications, sentiment analysis is a case in point where machine learning applications have led to significant advancements (Dhaoui *et al.*, 2017; Na and Thet, 2009). For example, expert application of machine learning based sentiment analysis provides insights for protecting and developing brands on social media against fans of rival brands (Ilhan *et al*., 2018), and machine learning models can automatically predict the helpfulness of online reviews in order to aid and enhance customers' online shopping experience (Singh *et al*., 2017).

Advances have also been made in different types of machine learning applications for marketing; *viz*. hybrid unsupervised machine learning approaches for improving customer lifetime value predictions (see: Hu *et al.*, 2013), and semi-supervised machine learning for fine-tuning marketing campaigns based on customer responses (and non-responses) (see: Lee *et al.*, 2010). In spite of these advancements, marketing literature still lacks appreciation of innovative methods developed and well-applied in other subject domains (Davis *et al*., 2013). Given this, there is room for further studies utilising machine learning methods in advancing marketing theory and practice (Balducci and Marinova, 2018).

In terms of content classification, researchers have shown how e-Word-of-Mouth (eWoM) can be auto-classified and mapped onto customer journeys for a better understanding of purchase decisions (Vázquez *et al*., 2014) and to aid in business engagement (Zhang *et al*., 2011), and recent work has demonstrated a machine learning based approach for classifying academic articles (Antons and Breidbach, 2018). Yet, a comparative approach to evaluate different machine learning techniques is somewhat rare; though Abu-Salih *et al*.'s (2018) classification of Twitter users' domain-specific interests is an important contribution in this context. Notwithstanding, compared to dealing with opinion valence in sentiment analysis, or short

snippets of textual user-generated content such as tweets, classifying news articles presents with additional challenges.

One challenge is that automatic classification of online content can be viewed as a multilabel classification problem, wherein the subject involves multiple tags/keywords (Salminen *et al.*, 2018). A 'label' is a machine learning term that refers to a sample in the training set. Multilabel classification is where an article, image etc. can be assigned multiple labels, such as when a feature film belongs to several different genres. Multilabel classification problems present themselves in a variety of contexts including, marketing messages on Twitter (Machedon *et al.*, 2013), toxic online comments (Salminen *et al.*, 2018), legal and economic articles (Mencía and Fürnkranz, 2008; Vogrinčič and Bosnić, 2011), and when classifying music into emotions (Trohidis *et al.*, 2011). However, acquiring training data for multilabel classification is not easy, as publicly available datasets are scarce and often limited in scope.

Another key challenge in the case of content marketing is that new topics emerge frequently due to emergence of new concepts and consumer interests, increasing the range of tags necessary to accurately capture the content collection. Moreover, algorithms are usually trained specifically on the type of content that they are subsequently applied to predict or classify, rather than considering the multichannel environment. For instance, an algorithm applied to website content is not necessarily expected to be effective across channels that vary in content-type, such as when classifying online videos whose titles and descriptions tend to be considerably scarcer than website content such as news and blog articles. As such, a machine learning model that is able deal with multilabel classification, frequent emergence of new topics, and is applicable across another channel with different type of data, would be of much value for marketers in terms of content optimisation and consistency of offerings across channels.

3. Overview of solution strategy

3.1 Algorithm selection and data cleaning

Many algorithms are not well-optimised for dealing with the problem at hand, since they do not possess the inbuilt capability of handling multilabel classifications. There are alternative methods to train multilabel classifiers, such as training one model for each label. However, since we are predicting news keywords, which are numerous and diverse, this approach is not technically feasible. As such, we have opted to evaluate three algorithms that have inbuilt

multilabel classification capabilities: Random Forest, K-Nearest Neighbors, and a Neural Network. In addition to being suitable for the problem at hand, these algorithms are publicly available in the Python programming language's Scikit-learn[2], a free software machine learning library that is widely applied and that we will also use in the modelling task[3].

The data cleaning process is essential in machine learning projects, particularly in this case since the raw data consists of the large strings, i.e., the news articles. These articles have noise such as text that is related to the website itself instead of the news story, words that occur commonly in articles, and the actual content of the article. The latter is the one in which we are interested. Therefore, we conduct data cleaning to eliminate irrelevant text content, including removing extra white space characters, non-alphabetic characters, and stopwords (i.e., words that have no actual meaning in the text like 'and', 'the', 'or', etc). After this, we utilise the well-known Term Frequency-Inverse Document Frequency (TF-IDF) algorithm (*Salton and Buckley, 1988*) to convert the cleaned article content into a numerical format, for easier consumption for the learning algorithms. Finally, we make use of cross-validation and parameter optimisation to obtain the best model and use it to predict news keywords for the articles in the data that are missing keywords.

3.2 Evaluation metrics

To evaluate the quality of the multilabel classification algorithm, a proper evaluation metric is needed. Since some keywords appear infrequently, we cannot use accuracy as a metric, as predicting no keywords most of the time will yield high accuracy. Therefore, a metric that takes both multiple labels and the frequency of keywords into account is needed. The *F1 Score*, which is the harmonic mean of two other metrics, Precision and Recall, is deemed suitable for this purpose (Wallach *et al.*, 2009). The F1 Score ranges between 0 and 1, where 1 is the ideal value. Equation 1 shows how this metric is calculated.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \qquad (1)$$

Precision measures how well the model avoids assigning the wrong keyword to an article; it is the number of true positives, positive instances that were classified correctly, divided by the sum of true positives and false positives (negative instances that were classified as positive).

[2] https://scikit-learn.org/stable/
[3] In addition to Scikit-learn, we will use Keras, another open source neural network library written in Python.

In contrast, *Recall* measures how well the model assigns keywords correctly to an article; it is the number of true positives divided by the sum of true positives and false negatives. The harmonic mean of these two metrics yields the F1 Score; thus, taking into account both how well the model avoids the wrong keywords and how well the model assigns the correct keywords.

Because one news article can contain several keywords that characterise its content, we apply multilabel classification. The number of labels (classes) assigned to an article is determined by a threshold value that is computed for each article/keyword combination. Because the distribution of keywords in the dataset is imbalanced, we use the *weighted* F1 Score that takes into account how often a label appears in the data and works better overall for evaluating multilabel classifications. In brief, the weighted F1 Score is the average F1 Score of each keyword weighted by its support (i.e., the number of true instances for each keyword).

4. Data exploration and preparation

4.1 Data collection and exploration

Al Jazeera is a global news and media organisation, headquartered in Doha, Qatar. The main website (aljazeera.com) attracts traffic from nearly 200 countries and regions and has had on average over 15 million visits in 2018, of which roughly 42% comes from search and another 44% is direct (SimilarWeb, 2018). We collected the data by scraping the content of Al Jazeera's main website that distributes news stories. The resulting dataset contains information about the article's content, its title, the date it was made and its keywords. The data contains 21,709 web pages, of which 13,058 have been classified by journalists and editors for news keywords. The remaining 8651 (39.8%) have not been classified, but using machine learning we are able to classify them. Overall, there are 799 different news keywords used by the journalists creating the content.

When a news article is extracted from the web, it contains some information that is not useful for the classification task, such as JavaScript functions, file routes, and source tags. Hence, a data cleaning procedure is needed to filter out such noise. Accordingly, first we eliminate the "SOURCE:" tag present in a large number of articles, followed by the initial tags such as "NEWS /". We then eliminated certain patterns that occurred with blank text (e.g. "';//]]]]"). Subsequently, we remove all JavaScript objects and functions, by eliminating all text between brackets. Finally, we filtered out unnecessary characters such as: extra white space, non-

alphabetic characters, stop-words, or words that have no actual meaning in the text (e.g. and, the, or). A specimen extract from the resulting cleaned dataset is presented in Appendix 1.

The structured data were queried for the top 10 keywords for 2017, which ranked 'eid' as the most frequent keyword (see Appendix 2 for results). We also utilised TF-IDF method to extracting more information from words based on their frequency of occurrence in a document; an example is provided in Appendix 3, which is an overview TF-IDF keywords on the topic of the Syrian war. Further, we examine the correlation between keywords to understand which keywords appear frequently together (see Appendix 4 for results); a list of highly correlated (r>0.5) keywords is generated, so that we may determine keywords that are a combination of individual words (e.g. *space* and *NASA*).

4.2 Data preparation

In our model, the text comprises of the headline and body text of the article content. Machine learning models only take numbers as input; as such, the unit of analysis in our case is a numerical representation of text (i.e. *vector*). Hence, to convert our articles from text to numbers, we opted for the TF-IDF method over simply counting the number of appearances of each unique word in each article. TF-IDF assigns scores to each word, based on how common they are in a specific article, and how uncommon they are across all articles (de Oliveira and da Rocha, 2006; Ramos, 2003). In order to create the TF-IDF matrix, the limits of the percentage of frequency of words need to be assigned first, which prevents words that are too rare or too frequent from being included in the matrix. Only the already classified articles are used for this purpose. The calculation for TF-IDF is shown in Equation 2.

$$w_{ij} = tf_{ij} \times log_2 \frac{N}{n} \qquad (2), \text{ where}$$

w_{ij} = weight of word *j* in article *i*

tfij = frequency of word j in article i

N = number of articles in the dataset

n = number of articles where word j is present at least once

From a descriptive overview of the keyword distribution (see Table 1 and Figure 1), we identify that there are several keywords with a very high frequency of occurrence. Most keywords appear less than 500 times in the whole dataset; however, there are a few keywords that appear

very frequently, skewing the distribution. Even after limiting keywords to those that appear less than 1000 times, a similar distribution pattern was revealed (see Appendix 5). Subsequently, by separating the top ten keywords (Figure 2) into easy-to-understand bins, we observe that 82% of the keywords appear in 20 articles or less (Table 2). This small number of article examples for each keyword may not be sufficient for the model and only add noise; therefore, these are removed. The resulting training dataset contains 13041 articles. Cleaned articles are then converted into a TF-IDF matrix. Finally, training data and labels are assigned using a tag-count matrix; thus, completing preparations for data modelling.

INSERT TABLE 1 HERE

INSERT FIGURE 1 HERE

INSERT FIGURE 2 HERE

INSERT TABLE 2 HERE

5. Data modelling

5.1 Classifier models and evaluation

As mentioned previously, the models we can use are limited to those that support multilabel classification efficiently; that is, to avoid using multiple One-vs-Rest classifiers to create the model. Using multiple One-vs-Rest classifiers is computationally inefficient, because this entails creating one model per keyword, then using all models during prediction time (Read *et al.*, 2011). This means training a large number of models, which will only increase in number when the number of keywords increases. Consequently, we consider three state-of-the-art machine learning classifiers, described as follows:

- **K-Nearest Neighbour:** Assigns points to the data, compares them using a distance metric, and assigns a classification based on the labels of the nearest points.
- **Random Forests:** Creates multiple decision trees, or statistical data structures that split the data according to criteria which divide the label best and averages them to create a more balanced prediction.
- **Neural Network:** Computes several matrix multiplications to approximate a function from its input to its output.

Though all of three of these support multilabel classification, for Random Forests and KNN, it is better to apply a dimension reduction technique to the data before training (Svetnik *et al.*, 2003). Neural Networks work better with high dimensional data, so for these models this step is not necessary. Subsequently, Principal Component Analysis (PCA) was chosen as the dimension reduction technique, which attempts to minimise the variance between the data in a higher dimension and its potential lower dimensions. Applying PCA to the TF-IDF matrix created previously, a new *sqrt*(n) dimensional matrix, where 'n' stands for the number of variables in the current, non-reduced matrix. We then use cross-validation[4], and the weighted F1 score to evaluate these models. For the Neural Network (NN), the *Keras* library to is used to create the neural network architecture. Also, a custom class is created to cross-validate and evaluate the neural network, since Keras does not support *scikit-learn* levels of cross-validation by default. In comparison to K-Nearest Neighbor (Average F1 Score: 0.577) and Random Forests (Average F1 Score: 0.458), the NN model outperforms the other two models (Average F1 Score: 0.627; Average Precision: 0.677; Average Recall: 0.612).

We also compute the algorithms' runtime (i.e., time taken to fit the model); in this comparison, KNN is the fastest, taking only 0.184 seconds to run on a test set of 10,000 articles, whereas RF takes 5.612 and NN 14.668 seconds on the same data. While the relative differences may seem large, the NN does not have a performance bottleneck in practical use. It can be trained on millions of articles in a matter of hours, if need be. For example, a linear estimation (NN's runtime grows linearly with the amount of data: $y = 0.0015x - 0.0491$, $R^2 = 1$) shows that training the NN model with one million articles would take approximately 25 minutes on the tested office hardware (a standard laptop with Intel Core i5 and 16GB of RAM memory). Again, we expect no performance bottleneck in practice, as organizations rarely produce this much content (even if an organization published, say, 100 new blog stories or news articles a day, it would take 27.4 years to produce a million pieces of content).

For further evaluation, different textual features are compared to assess the NN classifier's performance. The results, presented in Table 3, show that the highest performance can be obtained by including all the available textual features, including article title, description, and body-text.

---

[4] Cross-validation randomly divides the training data into *k* groups; each group (also called "fold") is then evaluated separately and the average of performance across all folds presented as the aggregate performance score of the model.

INSERT TABLE 3 HERE.

One further set of comparative evaluations is made using three different feature vector generators: Term Frequency (TF), TF-IDF, and Doc2Vec, an unsupervised algorithm to generate numerical vectors that represent text documents (Le and Mikolov, 2014; Papagiannopoulou and Tsoumakas, 2014). The results are summarised in Table 4.

INSERT TABLE 4 HERE.

Of the three feature vector types, TF-IDF performs the best. Therefore, we focus our efforts on optimising the parameters of this model. To do this, we create a helper class to perform random optimisation on both the TF-IDF matrix creation, and the Neural Network parameters. Subsequently, the best F1 Score for the combination of parameters is identified. With this, the model parameters are further fine-tuned using the grid search technique that experiments with different hyperparameters and chooses the combination that yields the best performance. Following this approach, *the final, optimised NN obtains the following performance scores*: Average F1 Score: 0.700; Average Precision: 0.685; Average Recall: 0.739. This also allows for the probability threshold that yields the highest F1 Score to established; the threshold value is 0.48, which means that a keyword is accepted by the model if it has a probability of more than 48%.

Finally, an important aspect of optimising a neural network model is determining the number of *epochs*. An epoch represents an iteration over the whole training set, such that the neural network updates the weights connecting each neuron. In our case, we observe that the optimal performance is obtained using four epochs (see Figure 3), after which the F1 Score begins to decrease.

INSERT FIGURE 3 HERE

5.2 Changes in keywords over time

The underlying structure of the data is highly likely to change with time given the volume and speed of data aggregation and the inherent nature of the data being aggregated in current data scenarios (e.g. with news content). This presents machine learning engineers with the issue known as *concept drift*, whereby the distribution of the underlying classification structure (e.g. labels) changes (Janardan and Mehta, 2017); for example, changes in keywords for new articles. As such, we assess our model's predictive performance also by training the NN with

only 2017 data and using 2018 data to test it. To explore how these keywords change from one year to the next, since our model will need to adapt to keywords changing over time, we visualise the relationship between 2017 and 2018 data (Figure 4).

INSERT FIGURE 4 HERE.

Some keywords appear much less frequently from one year to the next. This is because some topics (e.g. Puerto Rico from the hurricane in 2017), are not consistently relevant over time, in contrast to topics such as politics. To address this, our model checks for keyword counts in the past year and use only those that appear frequently during the whole year, with more emphasis on those that appear recently, as long as there is sufficient data on these. The F1 Score for 2018 data, which was trained only using 2017 data is: 0.625, which indicates that our model is able to perform acceptably on the new data, even though there is slight decrease of performance (10.7% decrease in performance compared to the optimized NN 2017 model). The result can be considered promising, especially given the large number of available classes for the neural network to tag the content. Generally, the probability of choosing the correct class by accident decreases with the increase in the number of classes, while the difficulty of finding the correct labels increases.

To investigate why the performance decreased when applying the model to "future" data, we conduct an LDA analysis (Latent Dirichlet Allocation), which is an unsupervised topic modelling algorithm (Blei *et* al., 2003; Li *et al*., 2018; Wong *et al.*, 2018), on the 2017 and 2018 datasets separately[5]. LDA is a Bayesian version of pLSA, i.e., Probabilistic Latent Semantic Analysis, that uses Dirichlet priors for the document-topic and word-topic distributions (Li *et al*., 2018; Newmann *et al*., 2011; Xu, 2018; Zhao *et al*. 2018). Using LDA, it is possible to infer human-interpretable topics from a text collection, such that each topic is characterized by the words that are most strongly associated with it. The results can be seen on Appendix 6, which shows how the 10 most prominent topics (retrieved as latent representations) change from one year to another. For example, Topic 1 is characterized by words such as "rohingya, iran, refugees, israel, like, women, and isil". We further compute the Jaccard coefficient that measures the overlap of two sets, which in our case are Set 1 = the unique topic keywords with the highest association to LDA-generated topics of the 2017

[5] We use the Gensim implementation of LDA in Python, available freely online: https://radimrehurek.com/gensim/models/ldamodel.html

dataset; and Set 2 = the unique topic keywords with the highest association to LDA-generated topics of the 2018 dataset. The calculation for the Jaccard coefficient is shown in Equation 3.

$J = \frac{N_c}{N_a + N_b - N_c}$ (3), where

N_a = Number of elements in Set 1

N_b = Number of elements in Set 2

N_c = Number of elements in the intersection of Set 1 and Set 2.

The Jaccard coefficient is 0.41, which indicates that the topics undergo considerable shift between 2017 and 2018. Due to the nature of our particular context, this is understandable – news topics change frequently according to real-world events. Given this, the obtained F1 Score of 0.625 can be considered as a fairly good result. We explain the relatively high F1 Score in this context as a consequence of the large body of training data and associated labels; because many keywords appear both in 2017 and 2018, the model is able to generalise from one year to another. However, if new keywords emerge in 2018 that are not present in the 2017 dataset, the model is unable to predict them at all.

6. Applying the model to predict keywords

6.1 Predicting keywords for news articles

As the first step in the process of predicting keywords, a total of 8160 articles missing their keywords were identified and converted into a TF-IDF matrix. Next, we use our trained model to predict which keyword(s) belong to each article. Since an article may have more than one keyword, the Neural Network computes a probability for each label to be present in an article; for selecting a label for an article, its probability must be ≥ 0.48. A specimen article, following keyword prediction, is provided in Appendix 7; the predicted keywords are intuitive and are contained within the article as well. Only 37 out of 8160 articles were left without keywords following prediction. The model was able to classify 8125 web pages out of 8161, yielding a success rate of 99.6%; here, the success rate is defined as the ability to classify confidently, where *confidence* is a threshold value of a model's internal accuracy.

In addition, by comparing the number of keywords given by the NN compared by the online content producers, we observe an interesting divergence. Whereas the online content producers were clearly biased in giving *three* labels per online content (see Figure 5), the NN model

applied a wider range when assigning the keywords, most often selecting 2–5 keywords ($\mu_{predicted}$ = 4.10 vs. μ_{real} = 3.54). When the content contains 5 or more keywords, the machine is substantially more efficient in findings matches (Figure 5). There are two implications for this – first, the online content creators' cognitive limits may decrease their ability to select relevant keywords (remember, the inventory of available keywords contains 799 possible choices). Second, if the former condition holds true, then the training data can be limited in its ability to describe the content pieces exhaustively, attributable to the fact that humans are simply not able to select all suitable keywords. This could result in an artificially low performance in the evaluation stage of the model because ground truth might be lacking some possibly matching keywords which the model would then assign to the content. However, even though these two claims are interesting, they are also speculative in nature; thus, future research should investigate the matter further.

INSERT FIGURE 5 HERE

6.2 Cross-channel evaluation using YouTube videos

Because of cross-platform content strategies (e.g. developing content for consumption across multiple platforms), a topic classification model developed for one channel could be deployed also to other channels. However, it is not necessarily the case that the model performs well in cross-platform deployment, mainly because the classification is based on text content and in different platforms the length and content descriptions vary. For example, a YouTube description is considerably shorter than a website article, therefore containing less information. To perform well in cross-platform deployment, the developed model needs to be able to deal with this fact of less (or more) information. Another challenge is that some of the content in one channel may have been tagged by the content creators manually, whereas tags are completely missing in another channel. From a machine learning point of view, this imposes a problem for the evaluation, since we are lacking ground truth (i.e., known keywords to evaluate the model against).

In order to evaluate how well the machine classification we trained with website content can be generalised across channels and content-type, we apply it to classify videos (titles and description) from the Al Jazeera's English language YouTube channel[6] , which has over 2.3

[6] www.youtube.com/channel/UCNye-wNBqNL5ZzHSJj3l8Bg

million subscribers at the time of writing. It should be noted that the lack of topics is also a concern for YouTube videos. The reason for this is that even though YouTube provides a way to categorise content, the available categories tend to be very general and thus, do not provide enough information to drive content marketing efforts. For example, in the case of Al Jazeera English, most content is classified under *News & Politics* on YouTube (see Figure 6 for an example), even though in our classification, the rubric of *News and Politics* has hundreds of sub-topics.

INSERT FIGURE 6 HERE

Overall, the model was able to classify 32,678 out of 33,996 of the YouTube videos, representing a success rate of 96.1%. This is on a par with the success rate of 99.6% obtained when classifying the website content. However, we have to evaluate the accuracy of these predicted labels in order to evaluate the real performance.

Since the model was not trained on the data that it is being used to classify, we have a lack of known values, or ground-truth for the cross-channel evaluation. As such, manual coding (i.e., human labelling) is needed to evaluate the performance of the model. We, therefore, employ three independent human coders, each rating the same 500 randomly sampled videos, assigning 1–3 labels per video. We then compare: (a) agreement between humans; and (b) agreement between humans and the machine (i.e., the optimised NN model). Because of the large number of available keywords, the probability of two raters choosing the same keyword by chance is small. For this reason, we use the simple percentage agreement between the raters as an evaluation metric. We calculate the agreement as follows:

a) For each row, there are nine possible different values, because each coder can choose three different classes (3x3).
b) For each row, we calculate agreement as [a = number of repeated values / 9]. A repeated value is the same value given by different coders. For example, if two coders label the item as “US politics”, and three coders as “Trump”, then the agreement is (2 + 3) / 9 = 0.56.
c) Finally, we average all items to get the overall simple agreement.

The results are summarised in Table 5, indicating that human coders agree with each other on the topics more than they do with the machine, but the difference is small, i.e., there is 10.4% higher agreement *between humans* than *between humans and neural network*. Overall, given the fact that even the human coders do not fully agree on the topics between them, and human-

to-machine agreement is very close to a human-to-human agreement, the model seems to generalise to a reasonable extent; i.e. it is able to assign meaningful topics to the videos, even though the model is trained on the website content that is much richer in terms of contained text than the average YouTube title and description. We attribute this successful result to the topical similarity between the organisation's content in the two channels; in other words, the content covers the same topics. Following this conjecture, we propose that the more overlapping the content between the organisations' various channels, the more likely a model developed using data from one channel is to generalise to other channels.

INSERT TABLE 5 HERE

7. Discussion and implications

There has been an increasing shift in the field of marketing from conventional forms of content analysis to more advanced computational forms corresponding to the vastly increasing availability, complexity, and importance of data (Balducci and Marinova, 2018; Cui *et al*., 2005; Kumar, 2018). Meanwhile, a parallel development in relation to research methodology in marketing has been called-for (Hofacker, 2012), so that innovative approaches may also contribute to greater advancements in marketing theory, especially by deriving deeper insights from unstructured, multi-faceted, and non-linear data (Syam and Sharma, 2018). Contributing towards this end, the current paper demonstrates an approach for taking unstructured online content, cleaning and structuring it for automatic tagging of multiple keywords by a Neural Network algorithm trained on already classified data from the website. We have also compared the performance of the Neural Network to two state-of-the-art multilabel classification algorithms, K-Nearest Neighbour and Random Forests, finding the Neural Network's performance to be better.

Although modern data-driven business scenarios can often be characterised by the abundance of large volumes of data (Kumar, 2018), preparing the data and structuring it be of actionable value to a business is challenging, not least because of the variability in the effectiveness of available machine learning solutions in comparison to the multitude of data problems (Flake *et al.*, 2004; Syam and Sharma, 2018). A comparative evaluation such as ours is therefore of value, especially to small and medium enterprises and start-ups, given the resource and time limitations for self-evaluation of available approaches (Abu-Salih *et al*., 2018). Unlike even some advanced clustering approaches used in marketing, for example, to classify text-based online reviews (Moon *et al.*, 2014), machine learning approaches such as what we have

demonstrated can be utilised for classifying full-length articles, and deal with multiple keywords per article.

Moreover, the application of the developed model to a different channel (YouTube) and content-type (video) has yielded promising results, reflecting positively on its generalisability. Cross-platform applicability of machine learning models is important because companies tend to be present in multiple social networks; e.g. publishing content on their website, Facebook, Instagram, YouTube, and so on. Often, the content they publish relates to similar topics, such as the case for a news and media organisation, because the topics are defined by the type of business and therefore, command content marketing efforts (Rowley, 2008). By automatically classifying the content across platforms, it becomes also possible to combine the data of various performance metrics into one aggregated analysis (e.g., analyse how audiences in different platforms responded to content on any specific topic). This presents opportunities not only for news agencies, but also for any other type of content creators including media and creative organisations, consultancies, academic journal publishers and research databases, to manage their content better and optimise it for the searching and sharing oriented digital consumer space. For instance, in the case of the focal organization that disseminates content in multiple online channels, the content in these channels is tagged sporadically, mainly due to the large number of content pieces produced, the lack of content marketing supervision, and the fact that multiple individuals with varying levels of expertise are involved in the tagging process. We surmise that this situation is common in the field and that most organisations are not efficiently tagging their online content. For such organisations, the introduction of an auto-tagging model is ideal.

Our approach can help to curate and seed content by desired criteria (e.g. customer interests), which is beneficial for firms that adopt content marketing as a business model or as part of a marketing strategy (Kilgour *et al*., 2015); and similarly, for researchers in accessing and understanding a vast corpus of research articles on a specific topic (Antons and Breidbach, 2018; Cates *et al*., 2017).

To better understand user intentions, motivations, and preferences, UGC and eWoM could also be classified using the same machine learning approach, since it offers an advantage over traditional statistical approaches for effectively dealing with large volumes of UGC that may manifest in different forms (e.g. text, videos) and across different platforms (cf. Abu-Salih *et al*., 2018; Uchinaka *et al.*, 2019). However, in that context attention needs to be paid to special

characteristics of UGC, e.g., bot content, humour, sarcasm and other noise factors in the data that can bias the classification. Effective curation of content in this way can ultimately lead to improvements in the overall business model, such as in developing better ranking systems for travel destinations (Ghose *et al.*, 2012). Furthermore, as our model's performance is acceptable when applied to future data in comparison to training data, news articles and other types of content on new and emergent topics could be classified automatically using the neural network approach we have demonstrated, similar to the way in which machine learning models have been applied for the classification of new products entering a market into existing categories of product types (Pandey *et al.*, 2018).

8. Limitations and suggestions for further research

One improvement to our study would be to obtain more data, more keywords, and more articles, to further expand and improve the capabilities of the model. Though a small number of articles remained unclassified (0.453% overall), to remedy this, we may either include more keywords during training, or decrease the probability threshold for accepting predicted keywords. However, both approaches have their disadvantages, including an increase in false positives due to lowering the threshold for keyword-acceptance. Another potential development is to explore more parameter and Neural Network architecture combinations, to obtain even better performance of the model. For example, 'bagging' Neural Networks could be applied in this respect to improve performance (Ha *et al.*, 2005).

Regarding the observed concept drift, we urge the organisations that deploy machine learning models to continuously track and monitor their performance and retrain them when the performance falls below a specified threshold. This threshold value is domain-specific and there are is no general value for F1 Score, for instance, that would apply in all domains. However, in the context of online content classification, an F1 score of 0.70 can be considered as satisfactory, as the vast majority of the content is correctly classified. When needed, the retraining of the models can be done through feeding the model more training data that captures the change in topics. Other possible techniques include combining labels to increase training data per class (assuming that the combined labels are conceptually associated) and exclusion of classes whose F1 scores fall below the defined threshold value.

Future research should address the question of implementing the developed model into practice in order to facilitate the content management and analysis process in the focal organisation, and to investigate its impact on the organization's workflow as well as evaluate the desired

efficiency gains and data quality improvements that the model is aiming to deliver. The provision of more personalised experiences for users by enabling adaptive web designs for example, is a pivotal in the viability of current business models and has consistently been called-for in the field of marketing (Kumar, 2018; Montgomery and Smith, 2008). Further studies may consider our results in this context and address questions about how effective and efficient classification of (unstructured) online content may improve navigability, accessibility, and share-ability of content; thus, eventually lead to the creation of customer 'value-in-use' of that content (Rowley, 2008). In addition, studies may examine how such automated classification may contribute to the effectiveness of marketing campaigns; for instance, meta-keywords are influential in search-engine marketing (Bing *et al.,* 2010), particularly by increasing search engines' ability to index websites correctly (Evans, 2007; Zhang and Dimitroff, 2005), but also in determining the efficacy of paid online search campaigns (Klapdor *et al*., 2014).

Finally, the current study only compared the text-based predictions to video content. Future research may expand the cross-channel applicability of our approach by attempting classification of other forms of online content such as images and video content, where machine learning has already been successfully applied in the marketing context (see: Deng and Li, 2018). Although our fully-supervised model was able to adapt to changes in the keywords over time, it does not have the ability to predict keywords that do not reoccur over time. As emerging research indicates (Zarrinkalam *et al.,* 2018), more dynamic changes in the content may be classified using unsupervised machine learning techniques that detect previously unknown patterns from the data. Nascent studies have shown that semi-supervised approaches, whereby a small part of the unstructured data is also used for training the model, can perform better, leading to cutting-edge applications in marketing (see: Ilhan *et al.,* 2018; Lee *et al.,* 2010).

9. Concluding remarks

Leveraging the benefits of machine learning applications in marketing and addressing the important need for such application for marketing research methods, this paper contributes to the literature by comparing three state-of-the art algorithms for tagging online website content and establishing cross-platform applicability. We find that the Neural Network performs the best for multilabel classification, and the developed model was able to cope with changes in

topics over time, which is salient in relation to news websites. Further, when applied to YouTube, the model provides an accuracy that is close to a human-to-human agreement.

References

Abu-Salih, B., Wongthongtham, P. & Chan, K. Y. (2018) 'Twitter mining for ontology-based domain discovery incorporating machine learning', *Journal of Knowledge Management,* 22 (5), pp. 949-981.

Antons, D. & Breidbach, C. F. (2018) 'Big data, big insights? Advancing service innovation and design with machine learning', *Journal of Service Research,* 21 (1), pp. 17-39.

Balducci, B. & Marinova, D. (2018) 'Unstructured data in marketing', *Journal of the Academy of Marketing Science,* 46 (4), pp. 557-590.

Bing, P., Zheng, X., Law, R. & Fesenmaier, D. R. (2010) 'The Dynamics of Search Engine Marketing for Tourist Destinations', *Journal of Travel Research,* 50 (4), pp. 365-377.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) 'Latent dirichlet allocation', *Journal of machine Learning research,* 3 (Jan), pp. 993-1022.

Cates, S., Lawrence, S., Penedo, C. & Samatova, V. (2017) 'A machine learning approach to research curation for investment process', *Journal of Investment Management,* 15 (1), pp. 39-49.

Chun, S.-H. (2018) 'Machine Learning Techniques and Statistical Methods for Business Applications: Implications on Big Data Gold Rush', *Advanced Science Letters,* 24 (7), pp. 5474-5477.

Clarke, T. B., & Jansen, B. J. (2017). Conversion potential: a metric for evaluating search engine advertising performance. *Journal of Research in Interactive Marketing*, 11(2), 142–159.

Cui, D. & Curry, D. (2005) 'Prediction in Marketing Using the Support Vector Machine', *Marketing Science,* 24 (4), pp. 595-615.

Cui, G. & Wong, M. L. (2004) 'Implementing Neural Networks for Decision Support in Direct Marketing', *International Journal of Market Research,* 46 (2), pp. 235-254.

Cui, G., Wong, M. L. & Lui, H.-K. (2006) 'Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming', *Management Science,* 52 (4), pp. 597-612.

Davis, D. F., Golicic, S. L., Boerstler, C. N., Choi, S. & Oh, H. (2013) 'Does marketing research suffer from methods myopia?', *Journal of Business Research,* 66 (9), pp. 1245-1250.

de Oliveira, R. & da Rocha, H. (2006) 'Mobile Access to Web Systems Using a Multi-device Interface Design Approach', *International Conference on Pervasive Systems & Computing*. Las Vegas, USA, pp. 37-46.

Deng, N. & Li, X. (2018) 'Feeling a destination through the "right" photos: A machine learning model for DMOs' photo selection', *Tourism Management,* 65, pp. 267-278.

Dhaoui, C., Webster, C. M. & Tan, L. P. (2017) 'Social media sentiment analysis: lexicon versus machine learning', *Journal of Consumer Marketing,* 34 (6), pp. 480-488.

Evans, M. P. (2007) ‘Analysing Google rankings through search engine optimization data’, *Internet Research*, 17(1), pp. 21–37.

Flake, G. W., Frasconi, P., Giles, C. L. & Maggini, M. (2004) 'Guest editorial: Machine learning for the Internet', *ACM Transactions on Internet Technology*, 4 (4), pp. 341-343.

Ha, K., Cho, S. & MacLachlan, D. (2005) 'Response models based on bagging neural networks', *Journal of Interactive Marketing*, 19 (1), pp. 17-30.

Hofacker, C. F. (2012) 'On Research Methods in Interactive Marketing', *Journal of Interactive Marketing*, 26 (1), pp. 1-3.

Hong, W., Zheng, X., Qi, J., Wang, W. & Weng, Y. (2018) 'Project Rank: An Internet Topic Evaluation Model Based on Latent Dirichlet Allocation', *2018 13th International Conference on Computer Science & Education (ICCSE)*. IEEE, pp. 1-4.

Hu, Y. H., Tsai, C. F., Hsu, Y. F. & Hung, C. S. (2013) 'A comparative study of hybrid machine learning techniques for customer lifetime value prediction', *Kybernetes*, 42 (3), pp. 357-370.

Ghose, A., Ipeirotis, P. G. & Beibei, L. (2012) 'Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content', *Marketing Science*, 31 (3), pp. 493-520.

Hong, W., Zheng, X., Qi, J., Wang, W. & Weng, Y. (2018) 'Project Rank: An Internet Topic Evaluation Model Based on Latent Dirichlet Allocation', *2018 13th International Conference on Computer Science & Education (ICCSE)*. IEEE, pp. 1-4.

Ilhan, B. E., Kübler, R. V. & Pauwels, K. H. (2018) 'Battle of the Brand Fans: Impact of Brand Attack and Defense on Social Media', *Journal of Interactive Marketing*, 43, pp. 33-51.

Janardan, M. & Mehta, S. (2017) 'Concept drift in Streaming Data Classification: Algorithms, Platforms and Issues', *Procedia Computer Science*, 122, pp. 804-811.

Kilgour, M., Sasser, S. L. & Larke, R. (2015) 'The social media transformation process: curating content into strategy', *Corporate Communications: An International Journal*, 20 (3), pp. 326-343.

Klapdor, S., Anderl, E. M., von Wangenheim, F. & Schumann, J. H. (2014) 'Finding the Right Words: The Influence of Keyword Characteristics on Performance of Paid Search Campaigns', *Journal of Interactive Marketing*, 28 (4), pp. 285-301.

Kumar, V. (2018) 'Transformative Marketing: The Next 20 Years', *Journal of Marketing*, 82 (4), pp. 1-12.

Kutlu, M., Elsayed, T. & Lease, M. (2018) 'Intelligent topic selection for low-cost information retrieval evaluation: A New perspective on deep vs. shallow judging', *Information Processing & Management*, 54 (1), pp. 37-59.

Le, Q. & Mikolov, T. (2014) 'Distributed representations of sentences and documents', *International conference on machine learning*. pp. 1188-1196.

Lee, H.-j., Shin, H., Hwang, S.-s., Cho, S. & MacLachlan, D. (2010) 'Semi-Supervised Response Modeling', *Journal of Interactive Marketing,* 24 (1), pp. 42-54.

Li, X., Zhang, A., Li, C., Ouyang, J. & Cai, Y. (2018) 'Exploring coherent topics by topic modeling with term weighting', *Information Processing & Management,* 54 (6), pp. 1345-1358.

Machedon, R., Rand, W. & Joshi, Y. (2013) *Automatic Crowdsourcing-Based Classification of Marketing Messaging on Twitter: 2013, International Conference on Social Computing*. 8-14 Sept. 2013.

Martínez-López, F. J. & Casillas, J. (2009) 'Marketing Intelligent Systems for consumer behaviour modelling by a descriptive induction approach based on Genetic Fuzzy Systems', *Industrial Marketing Management,* 38 (7), pp. 714-731.

Mencía, E. L. & Fürnkranz, J. (2008) *Efficient Pairwise Multilabel Classification for Large-Scale Problems in the Legal Domain*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Montgomery, A. L. & Smith, M. D. (2009) 'Prospects for Personalization on the Internet', *Journal of Interactive Marketing,* 23 (2), pp. 130-137.

Moon, S., Park, Y. & Kim, Y. S. (2014) 'The impact of text product reviews on sales', *European Journal of Marketing,* 48 (11/12), pp. 2176-2197.

Na, J.-C. & Thet, T. T. (2009) 'Effectiveness of web search results for genre and sentiment classification', *Journal of Information Science,* 35 (6), pp. 709-726.

Newman, D., Bonilla, E. V. & Buntine, W. (2011) 'Improving topic coherence with regularized topic models', *Advances in neural information processing systems*. pp. 496-504.

Orriols-Puig, A., Martínez-López, F. J., Casillas, J. & Lee, N. (2013) 'A soft-computing-based method for the automatic discovery of fuzzy rules in databases: Uses for academic research and management support in marketing', *Journal of Business Research,* 66 (9), pp. 1332-1337.

Pandey, S., Muthuraman, S. & Shrivastava, A. (2018) 'Data Classification Using Machine Learning Approach', in S., T., S., M., J., M., KC., L., A., J. & S., B. (eds.) *Intelligent Systems Technologies and Applications. ISTA 2017.* Cham: Springer, pp. 112-122.

Papagiannopoulou, E. & Tsoumakas, G. (2018) 'Local word vectors guiding keyphrase extraction', *Information Processing & Management,* 54 (6), pp. 888-902.

Pereira, R. B., Plastino, A., Zadrozny, B. & Merschmann, L. H. (2018) 'Correlation analysis of performance measures for multi-label classification', *Information Processing & Management,* 54 (3), pp. 359-369.

Ramos, J. (2003) *Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).*

Read, J., Pfahringer, B., Holmes, G. & Frank, E. (2011) 'Classifier chains for multi-label classification', *Machine learning,* 85 (3), p. 333.

Rowley, J. (2008) 'Understanding digital content marketing', *Journal of Marketing Management,* 24 (5-6), pp. 517-540.

Salminen, J., Almerekhi, H., Milenković, M., Jung, S., An, J., Kwak, H., & Jansen, B. J. (2018). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In Proceedings of The International AAAI Conference on Web and Social Media (ICWSM 2018). San Francisco, California, USA.

Salton, G. & Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management,* 24 (5), pp. 513-523.

SimilarWeb (2018). Retrieved from: www.similarweb.com. Accessed: 15 July 2018.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S. & Kumar Roy, P. (2017) 'Predicting the "helpfulness" of online consumer reviews', *Journal of Business Research,* 70, pp. 346-355.

Sterne, J. (2018) 'From programming to statistics to machine learning for marketing', *Applied Marketing Analytics,* 3 (4), pp. 298-305.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P. & Feuston, B. P. (2003) 'Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling', *Journal of Chemical Information and Computer Sciences,* 43 (6), pp. 1947-1958.

Syam, N. & Sharma, A. (2018) 'Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice', *Industrial Marketing Management,* 69, pp. 135-146.

Trohidis, K., Tsoumakas, G., Kalliris, G. & Vlahavas, I. (2011) 'Multi-label classification of music by emotion', *EURASIP Journal on Audio, Speech, and Music Processing*, 2011 (1), p. 4.

Uchinaka, S., Yoganathan, V. & Osburg, V.-S. (2019) 'Classifying residents' roles as online place-ambassadors', *Tourism Management,* 71, pp. 137-150.

Vázquez, S., Muñoz-García, Ó., Campanella, I., Poch, M., Fisas, B., Bel, N. & Andreu, G. (2014) 'A classification of user-generated content into consumer decision journey stages', *Neural Networks,* 58, pp. 68-81.

Vogrinčič, S. & Bosnić, Z. (2011) 'Ontology-based multi-label classification of economic articles', *Computer Science and Information Systems,* 8 (1), pp. 101-119.

Wallach, H. M., Murray, I., Salakhutdinov, R. & Mimno, D. (2009) *Evaluation methods for topic models: Proceedings of the 26th annual international conference on machine learning*. ACM.

Xu, J. (2018). Topic Modeling with LSA, PSLA, LDA & lda2Vec. Retrieved February 28, 2019, from https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

Zarrinkalam, F., Kahani, M. & Bagheri, E. (2018) 'Mining user interests over active topics on social networks', *Information Processing & Management,* 54 (2), pp. 339-357.

Zhang, J., & Dimitroff, A. (2005) ‘The impact of metadata implementation on webpage visibility in search engine results (Part II)’, *Information Processing & Management*, 41(3), pp. 691–715.

Zhang, M., Jansen, B. J., and Chowdhury, A. (2011) ‘Influence of Business Engagement in Online Word-of-mouth Communication on Twitter: A Path Analysis’, *Electronic Markets: The International Journal on Networked Business*, 21(3), 161-175.

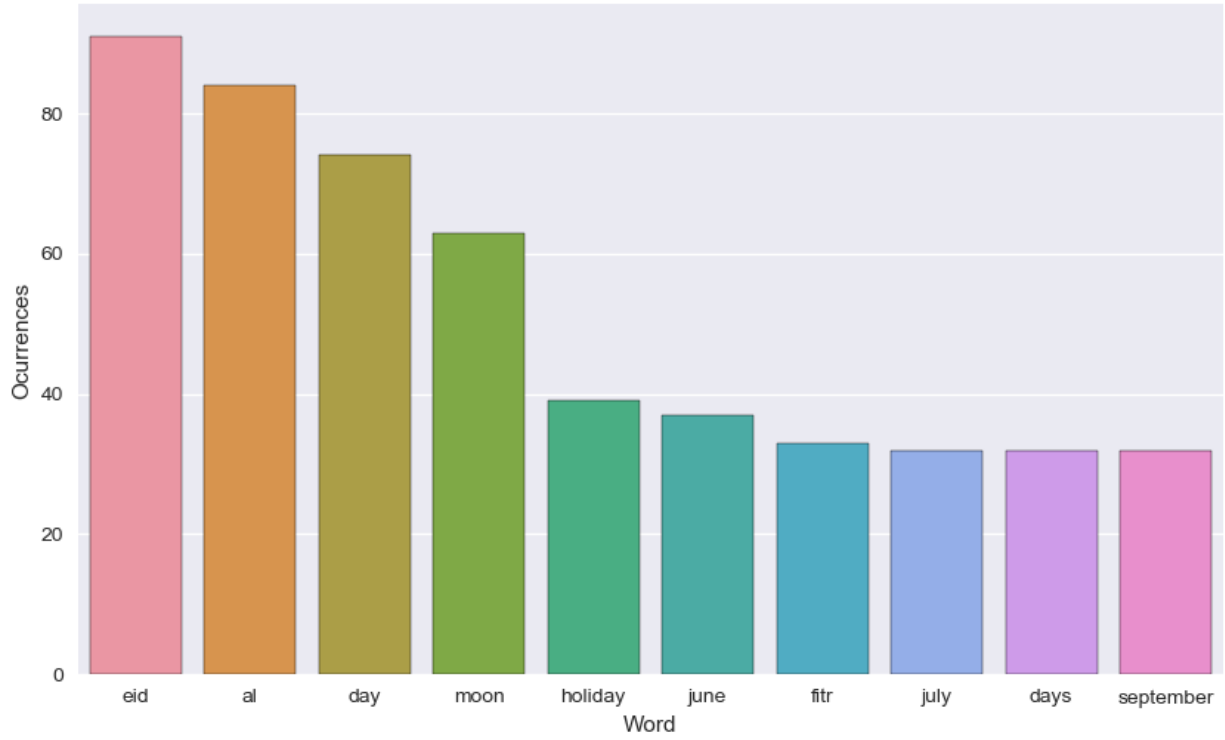
Zhao, W., Mao, J. & Lu, K. (2018) 'Ranking themes on co-word networks: Exploring the relationships among different metrics', *Information Processing & Management*, 54 (2), pp. 203-218.

Appendices

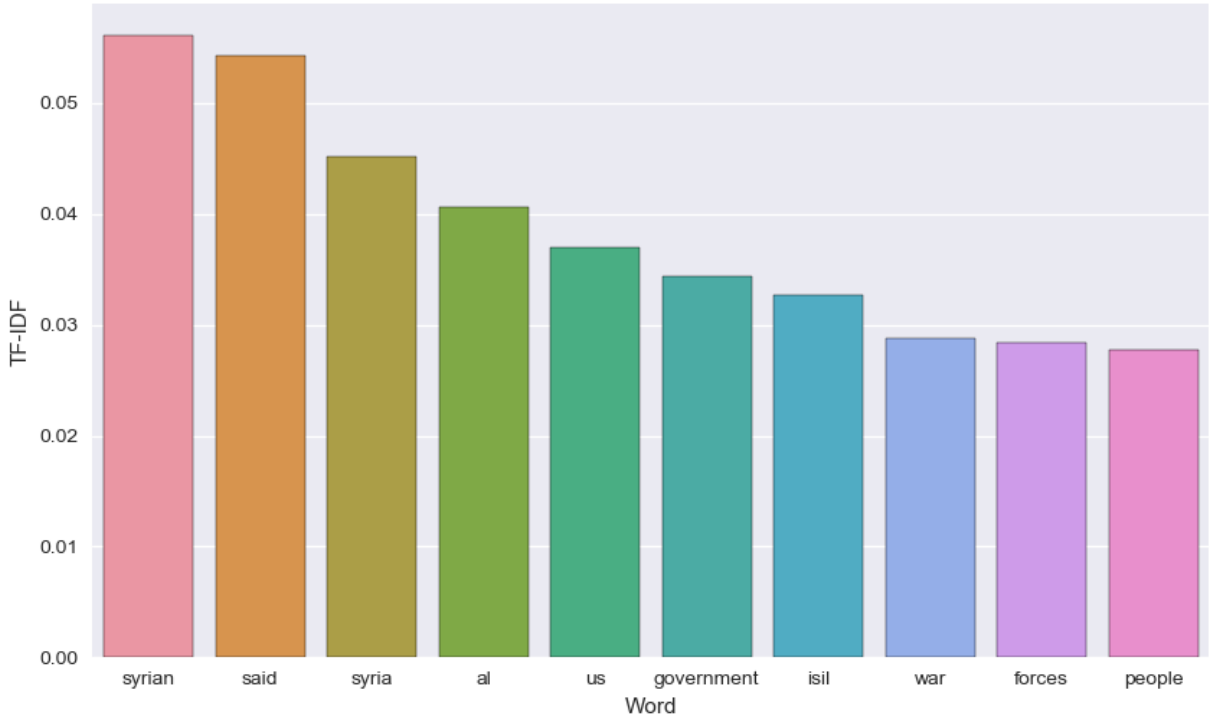
Appendix 1: Specimen article extract following data cleaning

Title	Description	Article body	News Keywords	Last modified date	Week	Month	Clean article body	predicted news keywords
0	The chronicler of Indian food	For 40 years, food historian Pushpesh Pant has...	Delhi, India - It is a winter afternoon and...	NaN	06/01/2016 11:21	1	1	chronicler indian food years food historian pu...
1	Netanyahu in India: What was swept under the c...		Narendra Modi Netanyahu in India: What was ...	NaN	NaT	-1	-1	netanyahu india swept carpet narendra modi net...
2	London Muslims Eid comes one week after attack	A van attack on worshippers in Finsbury Park -...	EuropeUK: Amber Rudd resigns in wake of Windru...	NaN	25/06/2017 19:40	25	6	london muslims eid comes one week attack van...
3	What's behind Hungary's campaign against Georg...		Politics Whats behind Hungarys campaign aga...	NaN	NaT	-1	-1	behind hungary campaign george soros politics ...
4	Venezuela: Mayhem rages; Capriles blocked from...		Venezuela Venezuelas Capriles says he was...	NaN	NaT	-1	-1	venezuela mayhem rages capriles blocked...

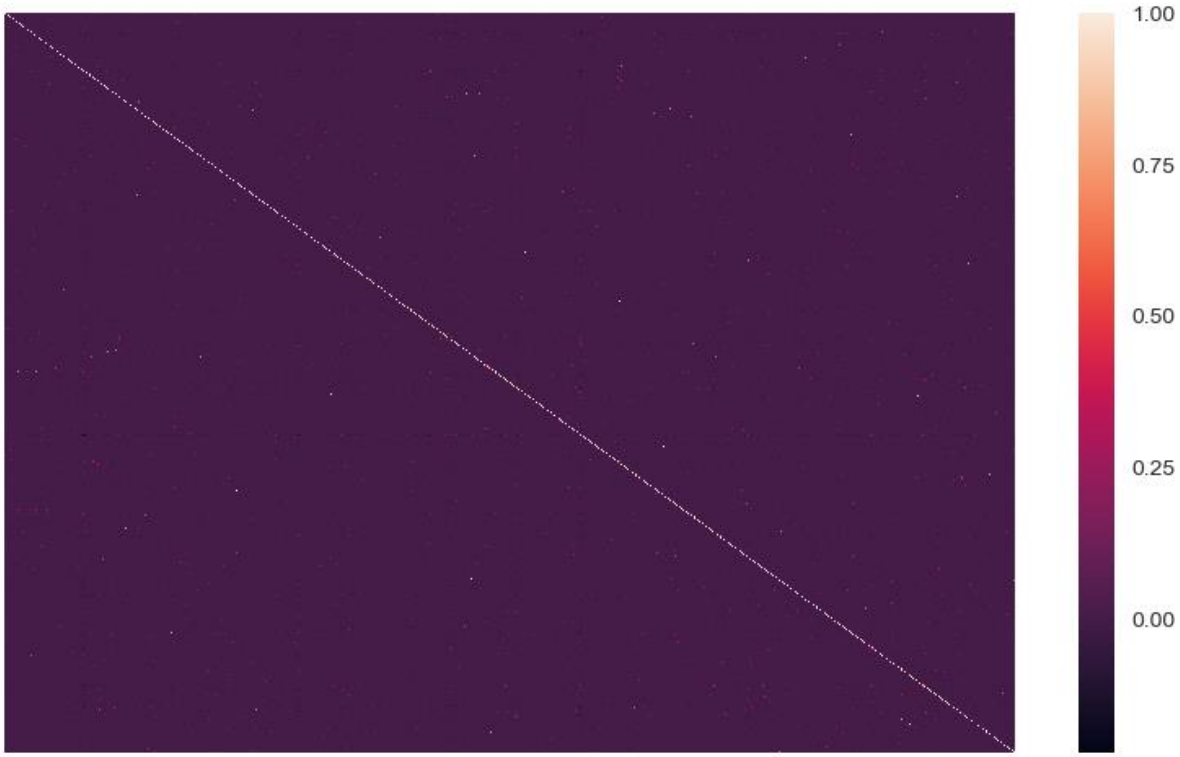
Appendix 2: Top 10 keywords in the dataset for 2017



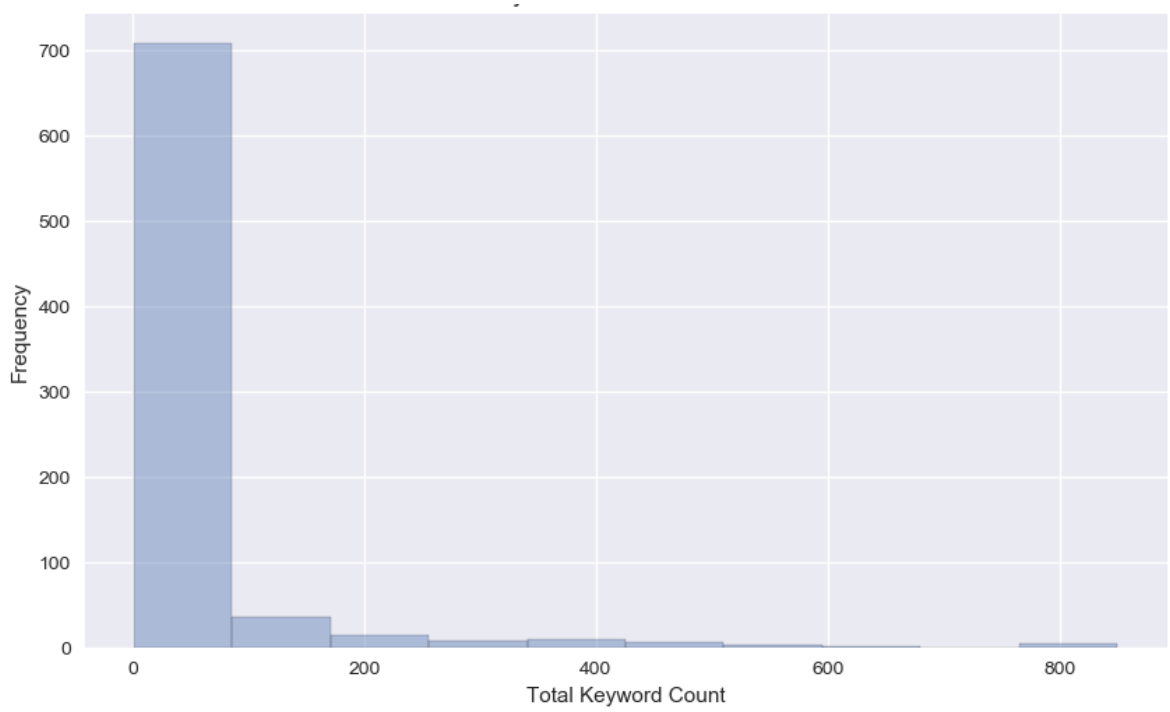
Appendix 3: TF-IDF keywords on the topic of the Syrian war



Appendix 4: Correlations between keywords



Appendix 5: Distribution of keywords appearing less than 1000 times



Appendix 6: Topics generated by LDA

Topics from 2017 as generated by LDA									
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
rohingya	israeli	myanmar	palestinian	bangladesh	palestinians	land	west	myanmars	bank
iran	deal	donald	trumps	house	washington	russia	white	nuclear	american
refugees	refugee	european	asylum	germany	family	europe	camp	border	children
israel	israeli	jerusalem	palestinian	palestinians	east	israels	palestine	arab	peace
like	university	white	even	history	life	black	way	see	never
women	muslim	ban	school	children	law	work	muslims	dont	like
isil	attack	fighters	suicide	afghanistan	afghanistans	attacks	afghan	taliban	bombing
qatar	gulf	saudi	arabia	countries	arab	uae	egypt	gcc	doha
percent	uk	economic	british	oil	africa	countries	economy	policy	business
function	india	pakistan	general	indian	muslim	muslims	indias	hate	court
Topics from 2018 as generated by LDA									
Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
women	children	like	school	life	family	old	even	dont	film
saudi	qatar	arabia	uae	yemen	bin	crisis	gulf	arab	prince
iran	donald	trumps	china	white	nuclear	house	american	washington	deal
percent	economic	change	water	economy	european	across	trade	countries	climate
attack	pakistan	least	taliban	children	function	afghan	armed	afghanistan	violence
police	court	muslim	anti	human	rohingya	law	arrested	groups	violence
syrian	syria	turkey	turkish	eastern	ghouta	kurdish	afrin	fighters	syrias
israeli	palestinian	israel	palestinians	gaza	jerusalem	israels	palestine	west	bank
south	north	korea	party	korean	opposition	election	elections	leader	talks
india	refugees	indian	refugee	russian	asylum	russia	uk	modi	prime

Appendix 7: Specimen article following keywords prediction

Predicted Keywords: europe, human rights, politics, refugees

"*Politics Whats behind Hungarys campaign against George Soros? Shrill campaign against man behind Open Society Foundations seen by critics as part of wider crackdown on civil society.by Patrick Strickland 22 Nov 2017 Rights groups and watchdogs say a Hungarian government campaign against investor and philanthropist George Soros has reached fever pitch, and it is being used to further a crackdown on civil society.Soros, an 86-year-old who was born in Hungary and is of Jewish descent, has been the focal point of attacks by Prime Minister Viktor Orban and his Fidesz party and other far-right nationalist outfits for years.Yet, recent months have seen a surge in anti-Soros rhetoric, that critics say is rooted in a desire to deflect attention from what they describe as a government crackdown on rights groups and civil society.Much of the antipathy stems from the policies advocated by the Open Society Foundations, a Soros-founded organisation that campaigns for strengthening civil society, advancing human rights and combating corruption.In Eastern and Central European countries, the Open Society Foundations has pushed for greater acceptance of refugees and migrants, putting it at odds with right-wing governments and far-right political parties.In July, the Hungarian government accused Soros of attempting to "Muslimise" Europe. Earlier this year, Orban, who is facing re-election in April 2018, led a campaign to shut down the Central European University (CEU), which was founded by Soros.On Monday, the Open Society Foundations pushed back, alleging in a statement that Orban and his political allies are orchestrating a campaign of "distortion and lies" about him, pointing to seven of Orbans statements that attacked Soros.Among those were claims that Soros hoped to resettle a million refugees in the European Union and allot them thousands of euros each.Balint Bardi, a Budapest-based Hungarian journalist, says the anti-Soros campaign is part of a broader strategy to "exploit the xenophobic feelings" of many Hungarians in order to "gain popularity for the government"."The government has been using this strategy since the beginning of the refugee crisis," Bardi told Al Jazeera by phone."They say there is a threat from our country from the migrants, from the politicians in Brussels or George Soros ... and that the government is the only one that can defend Hungarian society."He said the overwhelming focus on Soros compounds the anti-refugee propaganda and hostility towards international journalists and press outlets that do not support the government."This is very bad for Hungarian society," Bardi said.Attacking Hungary openlyAt a press conference on Monday, Gergely Gulyas, leader of the Fidesz parliamentary group, accused Soros of a "full frontal" attack on Hungary."So far, George Soros has attacked Hungary and the Hungarian government through the organisations he funds, the European Parliament and his allies in Brussels; but he has now entered the battle in person," Gulyas said, referring to the Open Society Foundations statement on Monday."George Soros is now attacking Hungary openly ... because in its immigration policy Hungary continues to stand its ground against the forces supporting immigration."Gulyas said Hungary "must not become an immigrant country".Contacted by Al Jazeera, the Hungarian governments International Communications Office declined to comment on the issue.The campaign against Soros has been unfolding alongside an apparent crackdown on civil society, including organisations affiliated with Soros and several that are not linked to him.In October, the Orban administration ordered the countrys intelligence services to investigate what it called an "empire" of Soros-backed institutions that work in Hungary.Nora Koves, a Hungarian human rights expert, said the government has increasingly targeted civil society institutions since 2013."Now its just continuing with Soros. Its not only the nongovernmental organisations (NGOs) being targeted and not only the migrants," said Koves, who works for the Budapest-*

based Eotvos Karoly Policy Institute, which has received funding from Soros-supported foundations."He is the perfect enemy because he is invisible and the Hungarian people will never meet him personally."In July, the parliament passed a law imposing strict rules on NGOs that receive foreign funding, requiring those that receive more than $26,000 a year from international sources to be registered as "foreign-supported".Last critics standingWith the strongest opposition groups being an increasingly fractious Socialist Party and Jobbik, an ultra-nationalist party accused of having neo-Nazi roots, Koves is holding out little hope for political pushback against the governments clampdown on civil society."Basically, we are the last critics standing in Hungary. The opposition is completely useless; people dont believe in them," she said. "But civil society is a whole different thing. We are the professional criticism of the government."They want to demolish it. If you want a perfect autocracy, then obviously you need to do this."Many in Hungary say the charges levelled at Soros, who survived the Holocaust, have an odour of latent anti-Semitism."The government is denying that is anti-Jewish propaganda against Soros, but many people think this is the case," Koves said.For years, governments across Central and Eastern Europe have blamed Soros for unrest and protests.Earlier this year, Romanias ruling party claimed that anti-corruption protests were orchestrated by Soros.In Poland, Jarosław Kaczynski, leader of the Law and Justice Party and a former prime minister, accused Soros-funded organisations of advocating "societies without identity".Anti-Soros measures and rhetoric have also become part and parcel of politics in countries including Serbia, Bulgaria and Slovakia.In the US and Europe, white supremacists and far-right commentators have pushed the widely debunked conspiracy theory that Soros was a Nazi collaborator, an officer in the German Schutzstaffel (SS) paramilitary and helped confiscate Jewish property for the Nazis and their allies during the second world war.Meanwhile, the Hungarian government, which has stridently opposed EU quotas on refugee distribution throughout member states, has styled itself as the defender of "Christian Europe" in the face of Muslim refugees, supposedly encouraged to come to Hungary by Soros and others.Lydia Gall, a Central and Eastern Europe researcher at Human Rights Watch, said that much of the anti-Soros rhetoric is "reminiscent of Nazi propaganda from the 1930s".Anti-Soros hoardingsGall alluded to government-funded anti-Soros hoardings visible across the capital and in small villages in the countryside, which often show images of Soros "depicted as the traditional grinning Jew" and play on "stereotypes that have been floating around against Jews for aeons of history"."The government is creating external enemies by linking refugees and asylum-seekers to terrorism, and claiming they are encouraged to come [to Hungary] by NGOs, which are in turn financed and supported by Soros," she told Al Jazeera.Referring to the anti-Soros tone of political discourse in Hungary, Serbia, Macedonia and Poland, among other countries, Gall said it should "prompt some action on behalf of the EU as a whole".In Hungary, she said, the strategy has been largely effective. An opinion poll published earlier this month found that the ruling Fidesz party maintains a 61-percent support rating, as reported by Hungarian Free Press. "When we see these types of illiberal and authoritarian tendencies in Europe and in the middle of the European Union, alarm bells should be ringing," Gall said."

Source: www.aljazeera.com

Tables and Figures

Table 1: Descriptives of keyword occurrences

Count		799.0
Mean		62.9
Std Deviation		256.5
Min		1.0
Max		4235.0
Quartiles	25%	2.0
	50%	8.0
	75%	28.0

Table 2: Top 10 keywords separated into bins

	Count	Cumulative Count	Cumulative %
(0, 10]	458	458	0.573
(20, 50]	103	561	0.702
(10, 20]	102	663	0.830
(50, 100]	51	714	0.894
(100, 250]	43	757	0.947
(250, 500]	22	779	0.975
(1000, 5000]	10	789	0.987
(500, 1000]	10	799	1.000

Table 3: Further evaluation of NN for different text features

Text Feature	**Mean F1 Score**	**Average Precision**	**Average Recall**
Title only	0.551	0.653	0.549
Title and description	0.426	0.658	0.588
Title, description, and body	0.627	0.666	0.643

Table 4: Further evaluation of NN for different feature vector generators

Feature vector generator	**Mean F1 Score**	**Average Precision**	**Average Recall**
TF	0.626	0.674	0.610
TF-IDF	0.640	0.667	0.642
Doc2Vec	0.516	0.571	0.514

Table 5: Comparing agreement for cross-channel results.

Metric	**%-agreement**	**Explanation**
h	77.7	Agreement among three human coders.
m1	69.6	Agreement between machine and humans, where the ratings of human coder 1 were replaced with machine ratings.
m2	71.0	Agreement between machine and humans, where the ratings of human coder 2 were replaced with machine ratings.
m3	70.6	Agreement between machine and humans, where the ratings of human coder 3 were replaced with machine ratings.
m_avg	70.4	Average agreement of replacing humans with machine ratings.

Notes: m1-m3 - each human coder was replaced by machine ratings in turn.

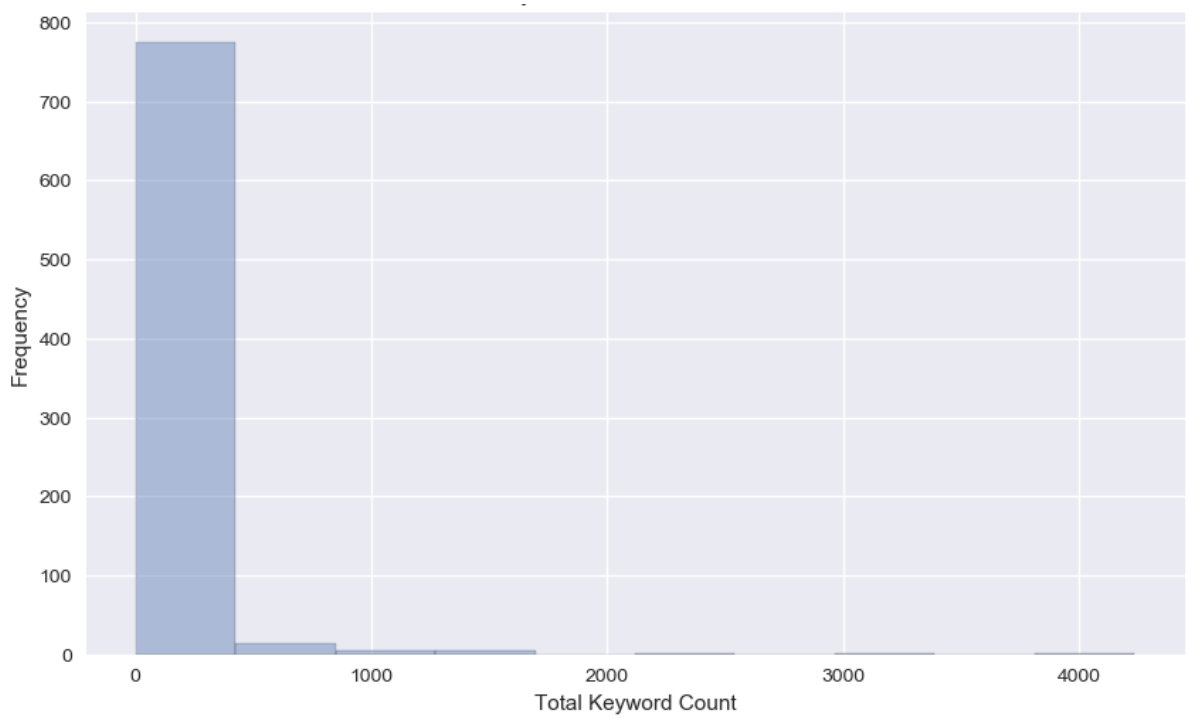


Figure 1: Distribution of keyword count with in the dataset

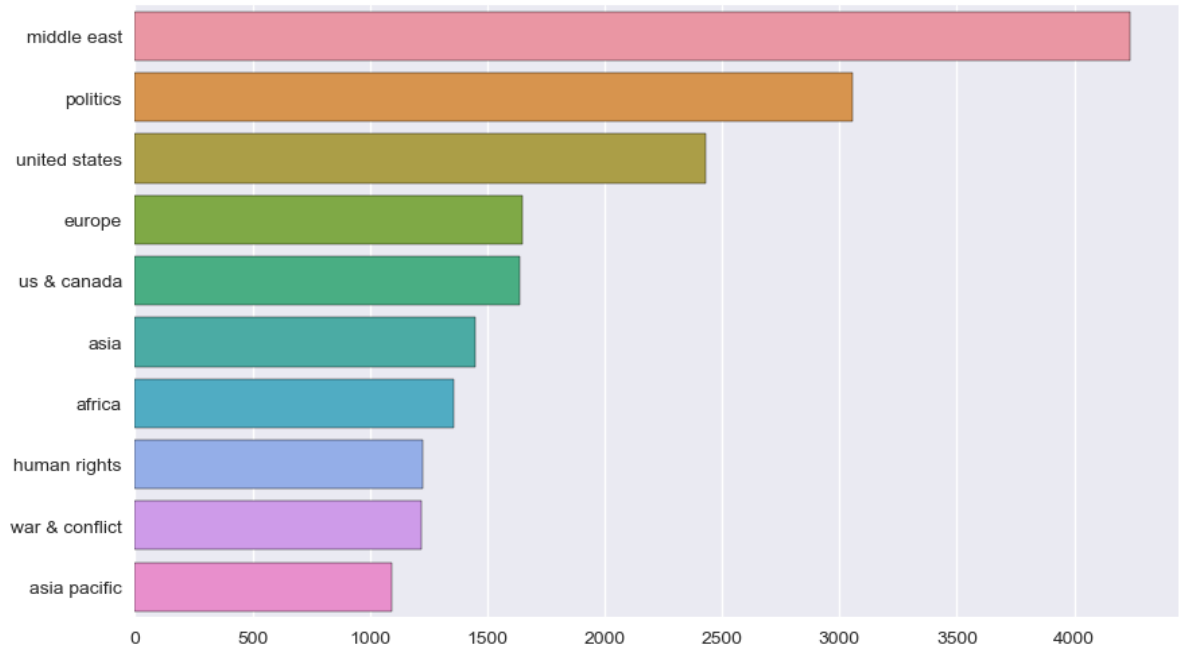


Figure 2: Top 10 keywords by frequency in the dataset

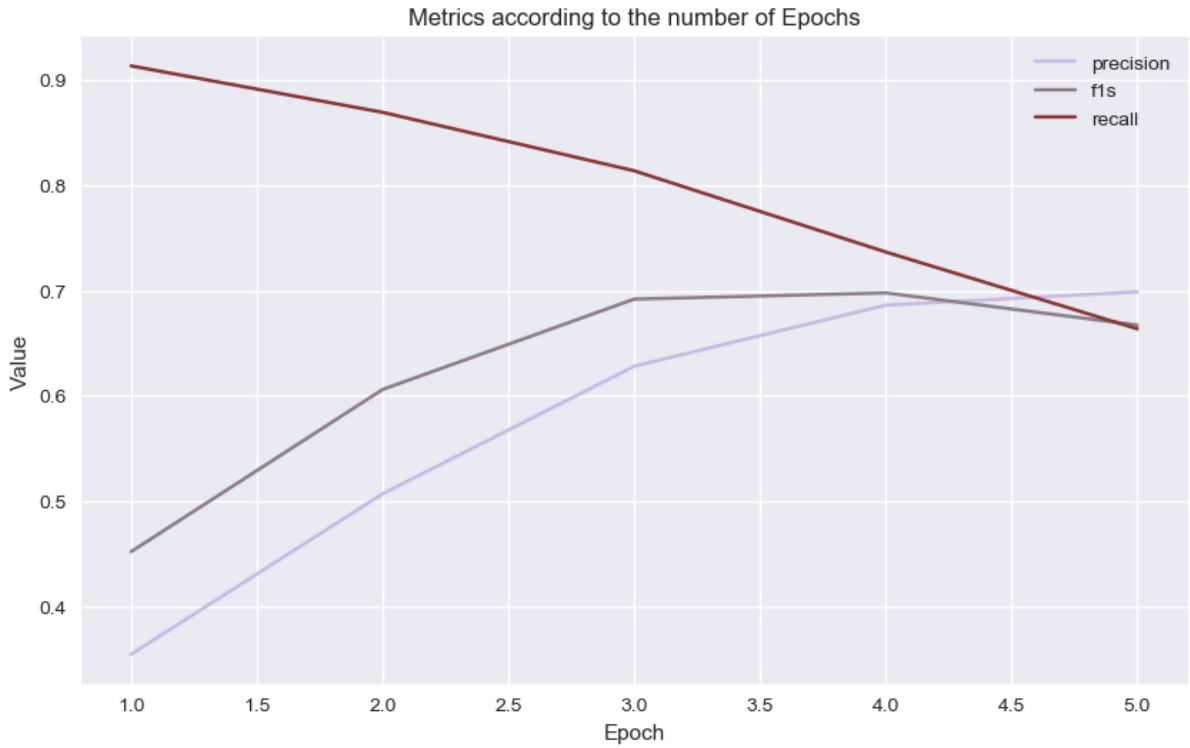


Figure 3: Performance metrics of the NN using different number of epochs

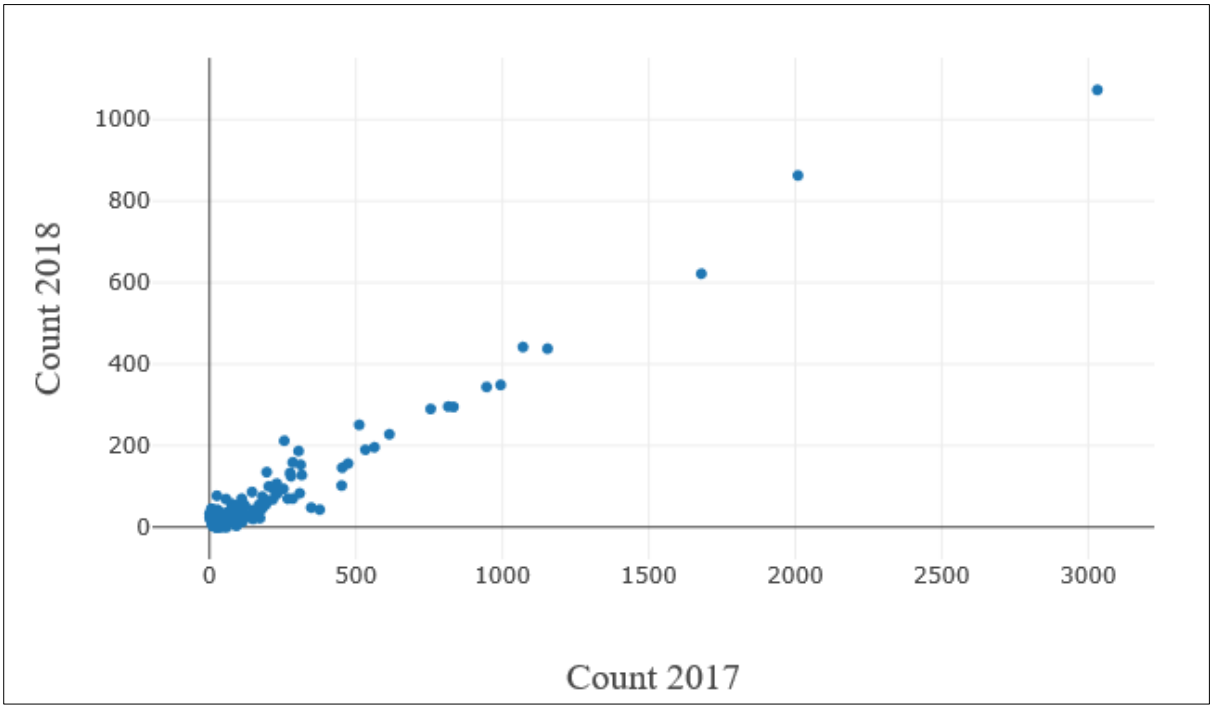


Figure 4: Evolution of keywords over a year

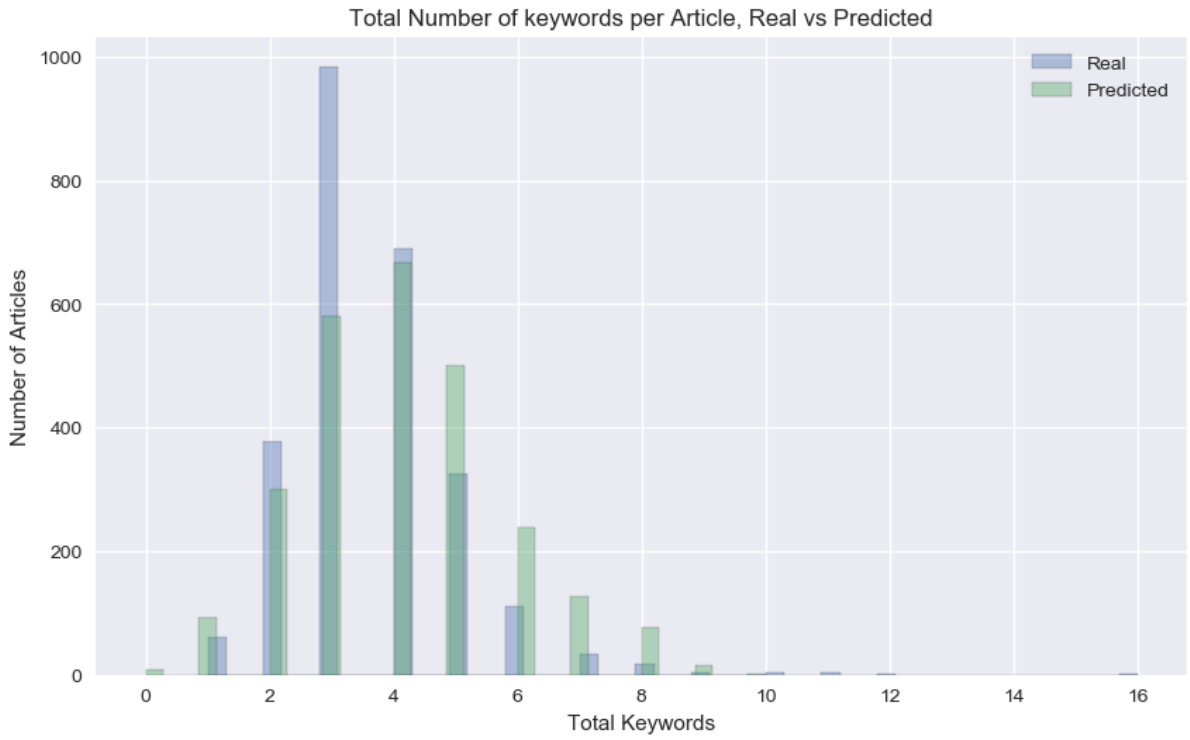


Figure 5: Number of keywords assigned by the online content creators ("Real") and the NN model ("Predicted")

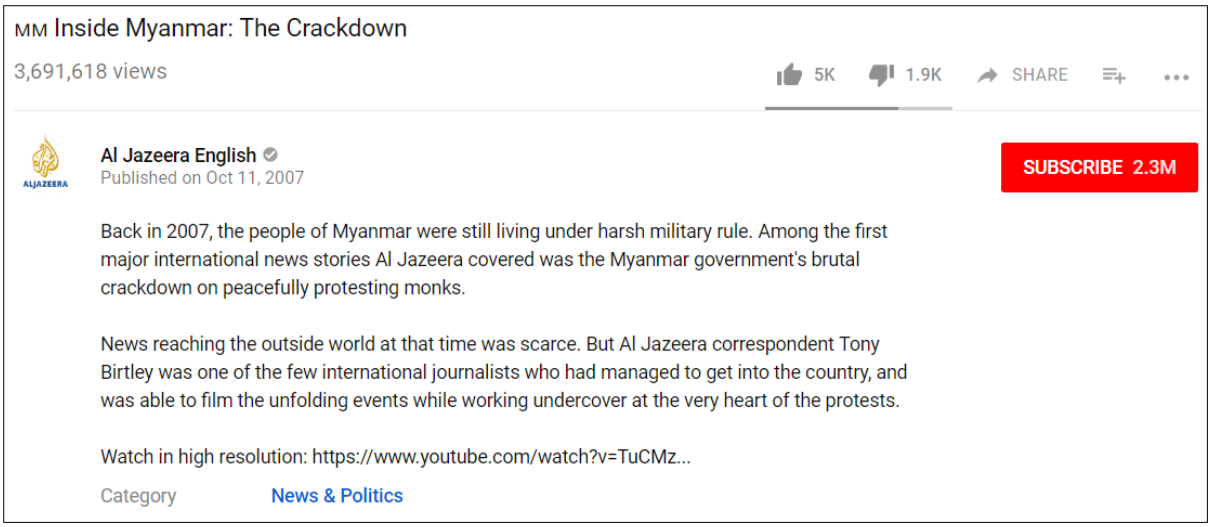


Figure 6: Example of YouTube categorisation under ‘News & Politics’